

## REMARKS

### Status of the Claims

Claims 1-7, 11-31, and 38-43 are pending in the present application. Claims 1, 13, and 23 have been amended to increase sequence identity to at least 90% and to recite functional language for fragments of the amino acid sequence set forth in SEQ ID NO:20. Support for these amendments may be found throughout the specification, for example on page 16, line 21, continuing through page 20, line 12. Claims 7 and 11 have been amended to correct improper articles, and claim 19 has been amended to add an article. No new matter has been added by amendment. Reexamination and reconsideration of the claims are respectfully requested.

The Examiner's comments in the Office Action are addressed below in the order set forth therein.

### The Objection to the Abstract Should Be Withdrawn

The abstract is objected to as it allegedly is not descriptive of the instant application. This objection is respectfully traversed. However, to advance prosecution, Applicants have amended the abstract to recite compositions and methods for orally active *Androctonus amoreuxi* pesticidal polypeptides. Applicants submit that the amended abstract describes the disclosure sufficiently to assist readers in deciding whether there is a need to consult the full text for additional details. Accordingly, reconsideration and withdrawal of the objection are respectfully requested.

### The Objection to the Title Should Be Withdrawn

The title is objected to as it allegedly is not descriptive of the instant application. This objection is respectfully traversed. However, to advance prosecution, Applicants have amended the title to recite orally active *Androctonus amoreuxi* pesticidal biopeptides. Applicants submit that the amended title is clearly indicative of the invention to which the claims are directed. Accordingly, reconsideration and withdrawal of the objection are respectfully requested.

The Objections to the Claims Should Be Withdrawn

Claims 7 and 11 are objected to for improper articles, while claim 19 is objected to for lack of an article. Claims 7 and 11 have been amended to correct the improper articles, and claim 19 has been amended to add an article. Accordingly, reconsideration and withdrawal of these objections are respectfully requested.

The Rejection of the Claims Under 35 U.S.C. § 112, First Paragraph (Written Description), Should Be Withdrawn

Claims 1-7, 13-19, 21-27, 29-38, 40, and 42 are rejected under 35 U.S.C. § 112, first paragraph, for allegedly failing to comply with the written description requirement of section 112. Specifically, the Office Action asserts that the claims contain subject matter which was not described in the specification in such a way as to reasonably convey to one skilled in the art that the inventors had possession of the claimed invention at the time the application was filed. This rejection is respectfully traversed as applied to claims 1-7, 13-19, 21-27, 29-38, 40, and 42.

Applicants have amended independent claims 1, 13, and 23 to recite a nucleotide sequence encoding a polypeptide having at least 90% sequence identity to the amino acid sequence set forth in SEQ ID NO:20, and a nucleotide sequence having at least 90% sequence identity to the coding sequence set forth in nucleotides 73-249 of SEQ ID NO:17 or nucleotides 64-240 of SEQ ID NO:14, wherein the nucleotide sequence encodes a polypeptide having pesticidal activity. Independent claims 1, 13, and 23 have also been amended to recite a nucleotide sequence encoding a polypeptide comprising a functional fragment of the amino acid sequence set forth in SEQ ID NO:20, wherein the polypeptide retains pesticidal activity.

The “Guidelines for Examination of Patent Applications Under the 35 U.S.C. 112, 1, ‘Written Description’ Requirement” state that a genus may be described by “sufficient description of a representative number of species . . . or by disclosure of relevant, identifying characteristics, i.e. structure or other physical and/or chemical properties.” 66 Fed. Reg. 1106 (January 5, 2001). This is in accordance with the standard for written description set forth in *Regents of the University of California v. Eli Lilly & Co.*, 119 F.3d 1559 (Fed. Cir. 1997), where the court held that “[a] written description of an invention involving a chemical genus, like a

description of a chemical species, 'requires a precise definition, such as by structure, formula, or chemical name' of the claimed subject matter sufficient to distinguish it from other materials." 119 F.3d at 1568, citing *Fiers v. Revel*, 984 F.2d 1164 (Fed. Cir. 1993).

The Federal Circuit has made it clear that sufficient written description requires simply the knowledge and level of skill in the art to permit one of skill to immediately envision the product claimed from the disclosure. *Purdue Pharma L.P. v. Faulding Inc.*, 230 F.3d 1320 1323, 56 USPQ2d 1481, 1483 (Fed. Cir. 2000) ("One skilled in the art must immediately discern the limitations at issue in the claims."). Amended claims 1, 13, and 23 recite a nucleotide sequence encoding a polypeptide having at least 90% sequence identity to the amino acid sequence set forth in SEQ ID NO:20, and a nucleotide sequence having at least 90% sequence identity to the coding sequence set forth in nucleotides 73-249 of SEQ ID NO:17 or nucleotides 64-240 of SEQ ID NO:14. The Office Action concluded that "the disclosure fails to describe the common attributes that identify members of the genus" and that the genus is "highly variant" (page 5, lines 10-11). However, the recitation of at least 90% sequence identity is a very predictable structural requirement encompassed by the claimed invention.

A satisfactory disclosure of a "representative number" of species depends on whether one of skill in the art would recognize that the applicants were in possession of the necessary common attributes or features of the elements possessed by the members of the genus in view of the species disclosed. 66 Fed. Reg. 1099, 1106 (2000). Applicants submit that the knowledge and level of skill in the art would allow a person of ordinary skill to envision the claimed invention, *i.e.*, a nucleotide sequence encoding a polypeptide having at least 90% sequence identity to the amino acid sequence set forth in SEQ ID NO:20, and a nucleotide sequence having at least 90% sequence identity to the coding sequence set forth in nucleotides 73-249 of SEQ ID NO:17 or nucleotides 64-240 of SEQ ID NO:14.

Furthermore, as described above, the description of a claimed genus can be by structure, formula, chemical name, or physical properties. See also *Ex parte Maizel*, 27 USPQ2d 1662, 1669 (B.P.A.I. 1992), citing *Amgen, Inc. v. Chugai Pharmaceutical Co., Ltd.*, 927 F.2d 1200, 1206 (Fed. Cir. 1991). The recitation of a predictable structure of a sequence comprising a nucleotide sequence encoding a polypeptide having at least 90% sequence identity to the amino

acid sequence set forth in SEQ ID NO:20, or a nucleotide sequence having at least 90% sequence identity to the coding sequence set forth in nucleotides 73-249 of SEQ ID NO:17 or nucleotides 64-240 of SEQ ID NO:14 is sufficient to satisfy the written description requirement. These structural limitations are sufficient to distinguish the claimed nucleotide sequences from other materials and thus sufficiently define the claimed genus.

Applicants have further provided the functional characteristic that distinguishes the claimed nucleotide sequences (*i.e.*, those encoding polypeptides having at least 90% sequence identity to the amino acid sequence set forth in SEQ ID NO:20, or those having at least 90% sequence identity to the coding sequence set forth in nucleotides 73-249 of SEQ ID NO:17 or nucleotides 64-240 of SEQ ID NO:14) of the genus. Specifically, the claims recite that the nucleotide sequences encode polypeptides having pesticidal activity; thereby providing a functional characterization of the nucleotide sequences claimed in the genus. Likewise, Applicants have provided the same functional characteristic for nucleotide sequences encoding a polypeptide comprising a functional fragment of the amino acid sequence set forth in SEQ ID NO:20. Accordingly, both the structural properties **and** the functional properties that characterize the claimed genus are specifically recited in the claims. Therefore, Applicants have conveyed with reasonable clarity to one skilled in the art that they were in possession of the claimed invention. Applicants show possession of the claimed invention by describing the claimed invention with all of its limitations using such descriptive means as words, structures, figures, diagrams, and formulas that fully set forth the claimed invention. *Lockwood v. American Airlines, Inc.*, 107 F.3d 1565, 1572, 41 USPQ2d 1961, 1966 (Fed. Cir. 1997).

In view of the above amendments and remarks, Applicants submit that all grounds for rejection under 35 U.S.C. § 112, first paragraph (written description), have been overcome. Accordingly, reconsideration and withdrawal of the rejection are respectfully requested.

The Rejection of the Claims Under 35 U.S.C. § 112, First Paragraph (Enablement), Should Be Withdrawn

Claims 1-7, 13-19, 21-27, 29-31, 38, 40, and 42 are rejected under 35 U.S.C. § 112, first paragraph, for allegedly failing to comply with the enablement requirement of section 112.

Specifically, the Office Action concedes that the specification is enabling for nucleic acids encoding SEQ ID NO:20, as well as for expression cassettes, host cells, viruses, plants, and seeds comprising these nucleic acids, and for methods of using the same. However, the specification “fails to provide guidance for how to make or where to find nucleic acids encoding pesticidal proteins with 80% identity to SEQ ID NO:20 or pesticide-encoding nucleic acids with 80% identity to bases 73-249 of SEQ ID NO:17 or bases 64-240 of SEQ ID NO:14. The instant specification also fails to provide guidance for how to use nucleic acids encoding 10 contiguous amino acids of SEQ ID NO:20 and nucleic acids comprising 30 contiguous nucleotides of bases 73-249 of SEQ ID NO:17 or bases 64-240 of SEQ ID NO:14” (Office Action, page 7, lines 16-21). This rejection is respectfully traversed as applied to claims 1-7, 13-19, 21-27, 29-31, 38, 40, and 42.

As discussed above, Applicants have amended independent claims 1, 13, and 23 to recite a nucleotide sequence encoding a polypeptide having at least 90% sequence identity to the amino acid sequence set forth in SEQ ID NO:20, and a nucleotide sequence having at least 90% sequence identity to the coding sequence set forth in nucleotides 73-249 of SEQ ID NO:17 or nucleotides 64-240 of SEQ ID NO:14, wherein the nucleotide sequence encodes a polypeptide having pesticidal activity. Independent claims 1, 13, and 23 have also been amended to recite a nucleotide sequence encoding a polypeptide comprising a functional fragment of the amino acid sequence set forth in SEQ ID NO:20, wherein the polypeptide retains pesticidal activity. In contrast to the conclusion reached in the Office Action, Applicants contend that support is provided for both the sequence identity limitations of the claims and the functional limitation of the claims. Guidance for determining percent identity of sequences is provided in the specification on page 16, line 21, continuing through page 18, line 5; and on page 23, line 14, continuing through page 28, line 18. The procedures for making nucleotide sequences encoding variants (*e.g.*, of SEQ ID NO:20) are conventional in the art (see, *e.g.*, page 18, lines 6-18 of the specification, which lists a number of exemplary references for such procedures). Additionally, the specification provides ample support for assays that are used to identify variants having the claimed pesticidal activity (see, *e.g.*, Example 5, *Corn Rootworm Bioassay*, and Example 6, *Homopteran Bioassay*). Thus, support is provided to enable one of skill in the art to make and

•

use a nucleic acid and/or nucleotide construct meeting the sequence identity limitations of the claims.

When rejecting a claim under the enablement requirement of section 112, the Examiner has the initial burden to establish a reasonable basis to question the enablement provided for the claimed invention. MPEP § 2164.04, citing *In re Wright*, 999 F.2d 1557, 1562, 27 USPQ2d 1510, 1513 (Fed. Cir. 1993). In this case, the Examiner relies heavily on “unpredictability, and lack of guidance” (Office Action, page 8, line 18). For example, on page 8, lines 18-19, the Office Action states that “[g]iven the claim breath, unpredictability, and lack of guidance ..., undue experimentation would have been required by one of skill in the art.” The Office Action cites Lazar *et al.* (*Mol. Cell. Biol.* 8:1247-52, 1988) and Hill and Preiss (*Biochem. Biophys. Res. Comm.* 244:573-77, 1998) for the proposition that “making substitutions is not predictable” (*i.e.*, because “conservative” amino acid substitutions affected protein biological activity; see page 8, lines 8-17 of the Office Action). However, Applicants respectfully contend that these isolated references simply illustrate that one of skill in the art would readily be able to determine whether a particular amino acid change affected a biological activity of a protein. Furthermore, as shown in attached **Exhibit A** (Bowie *et al.*, *Science* 247:1306-10, 1990), numerous evolutionary and mutagenesis studies have shown that proteins are highly plastic in tolerating amino acid changes. Consequently, one of skill in the art would be able to determine the functionality of polypeptides encompassed by the claimed nucleotide sequences without resorting to undue experimentation.

The Federal Circuit has repeatedly stated that enablement is not precluded by the necessity for some experimentation, so long as the experimentation needed to practice the invention is not undue, and that a considerable amount of experimentation is permissible if it is merely routine or if the specification provides a reasonable amount of guidance as to how the experimentation should proceed. *In re Wands*, 858 F.2d 731, 737, 8 USPQ2d 1400, 1404 (Fed. Cir. 1988). In the instant case, the quantity of experimentation required to practice the claimed invention amounts to two steps. First, generating a nucleotide sequence encoding a polypeptide having at least 90% sequence identity to the amino acid sequence set forth in SEQ ID NO:20, a nucleotide sequence having at least 90% sequence identity to the coding sequence set forth in nucleotides 73-249 of SEQ ID NO:17 or nucleotides 64-240 of SEQ ID NO:14, or a nucleotide

sequence encoding a polypeptide comprising a functional fragment of the amino acid sequence set forth in SEQ ID NO:20. Second, assaying the encoded polypeptide for functional activity. Such assays, while known in the art, have further been presented in the specification (see, *e.g.*, Examples 5 and 6). One of skill in the art would appreciate that both of these steps are within the skill of those in the art and that this degree of experimentation is not considered undue. “[A] specification disclosure which contains a teaching of the manner and process of making and using the invention in terms which correspond in scope to those used in describing and defining the subject matter sought to be patented *must* be taken as in compliance with the enabling requirement of the first paragraph of § 112 *unless* there is reason to doubt the objective truth of the statements contained therein which must be relied on for enabling support.” *In re Marzocchi*, 439 F.2d 220, 223, 169 USPQ 367, 369 (CCPA 1971) (emphasis in original).

Furthermore, Applicants contend that the Examiner’s statement that the specification “fails to provide guidance for how to use ... nucleic acids comprising 30 contiguous nucleotides of bases 73-249 of SEQ ID NO:17 or bases 64-240 of SEQ ID NO:14” (Office Action, page 7, lines 19-21) is incorrect. On page 14, line 23, continuing through page 16, line 20 of the specification is found clear guidance for use of fragments and variants of the disclosed nucleotide sequences, including, for example, use as hybridization probes or PCR primers. These and additional uses for fragments of isolated nucleotide sequences are well known to one of ordinary skill in the art.

In view of the above amendments and remarks, Applicants submit that all grounds for rejection under 35 U.S.C. § 112, first paragraph (enablement), have been overcome. Accordingly, reconsideration and withdrawal of the rejection are respectfully requested.

The Rejection of the Claims Under 35 U.S.C. § 112, Second Paragraph, Should Be Withdrawn

Claim 19 is rejected under 35 U.S.C. § 112, second paragraph, for allegedly being indefinite, and claims 23-31, 42, and 43 are rejected under 35 U.S.C. § 112, second paragraph, for allegedly omitting essential steps. This rejection is respectfully traversed as applied to claims 19, 23-31, 42, and 43.

The Office Action asserts that is unclear what type of expression construct was used to transform the seed of claim 19. Claim 19 depends from claim 13, which clearly recites a transformed plant comprising in its **genome** at least one **stably incorporated** expression cassette comprising a sequence of the invention. “By ‘stable transformation’ is intended that the nucleotide construct introduced into a plant integrates into the genome of the plant and is capable of being inherited by progeny thereof” (see page 37, lines 18-20 of the specification). One of skill in the art would clearly understand that the seed of claim 19 is transformed with the expression cassette stably incorporated into the genome of the parent plant. Accordingly, reconsideration and withdrawal of the rejection as applied to this claim are respectfully requested.

The Office Action further asserts that claims 23-31, 42, and 43 omit steps involved in regenerating a plant from a plant cell. Claims are considered to be definite, as required by the second paragraph of 35 U.S.C. § 112, when they define the metes and bounds of a claimed invention with a reasonable degree of precision and particularity (see, *e.g.*, *In re Venezia*, 530 F.2d 956, 958, 189 USPQ 149, 151 (CCPA, 1976). In the instant case, the Examiner has failed to give the rationale for considering the omitted steps **critical** or **essential** (see MPEP § 706.03(d), *Rejections Under 35 U.S.C. 112, Second Paragraph*, Examiner Note 3 following form paragraph 7.34.12). Furthermore, as stated in *Bendix Corp. v. United States*, “it is **not** necessary that a claim recite each and every element needed for the practical utilization of the claimed subject matter” (600 F.2d 1364, 1369, 204 USPQ 617, 621 (Ct. Cl. 1979)) (emphasis added). Applicants contend that they have properly defined the metes and bounds of the claimed invention as found in claims 23-31, 42, and 43 with a reasonable degree of precision and particularity. Accordingly, reconsideration and withdrawal of the rejection as applied to these claims are respectfully requested.

#### Prior Art and Allowable Subject Matter

Applicants thank the Examiner for noting that claims 1-7, 11-31, and 38-43 are free of the prior art. Applicants also thank the Examiner for indicating that claims 11, 12, 20, 28, 39,



Appl. No.: 10/617,978  
Amdt. dated January 5, 2006  
Reply to Office Action of October 7, 2005

41, and 43 would be allowable if rewritten in independent form. Applicants note that claims 11 and 12 are independent claims, and therefore respectfully suggest that these claims are allowable.

### CONCLUSION

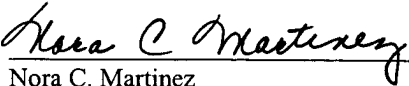
In view of the foregoing amendments and remarks, Applicants respectfully submit that all the objections and rejections have been obviated or overcome and the claims are in condition for allowance. Early notice to this effect is solicited. If, in the opinion of the Examiner, a telephone conference would expedite the prosecution of the subject Application, the Examiner is invited to call the undersigned attorney.

It is not believed that extensions of time or fees for net addition of claims are required, beyond those that may otherwise be provided for in documents accompanying this paper. However, in the event that additional extensions of time are necessary to allow consideration of this paper, such extensions are hereby petitioned under 37 CFR § 1.136(a), and any fee required therefore (including fees for net addition of claims) is hereby authorized to be charged to Deposit Account No. 16-0605.

Respectfully submitted,



David E. Cash  
Registration No. 52,706

<p><b>Customer No. 29122</b> <b>ALSTON &amp; BIRD LLP</b> Bank of America Plaza 101 South Tryon Street, Suite 4000 Charlotte, NC 28280-4000 Tel Raleigh Office (919) 862-2200 Fax Raleigh Office (919) 862-2260</p>	<p><b><u>CERTIFICATE OF EXPRESS MAILING</u></b> "Express Mail" mailing label number EV395777334US Date of Deposit: January 5, 2006 I hereby certify that this paper or fee is being deposited with the United States Postal Service "Express Mail Post Office to Addressee" service under 37 CFR 1.10 on the date indicated above and is addressed to: Mail Stop Amendment, Commissioner for Patents, P.O. Box 1450, Alexandria, VA 22313-1450  Nora C. Martinez</p>
---	--

# Deciphering the Message in Protein Sequences: Tolerance to Amino Acid Substitutions

JAMES U. BOWIE,\* JOHN F. REIDHAAR-OLSON, WENDELL A. LIM,  
ROBERT T. SAUER

An amino acid sequence encodes a message that determines the shape and function of a protein. This message is highly degenerate in that many different sequences can code for proteins with essentially the same structure and activity. Comparison of different sequences with similar messages can reveal key features of the code and improve understanding of how a protein folds and how it performs its function.

THE GENOME IS MANIFEST LARGELY IN THE SET OF PROTEINS that it encodes. It is the ability of these proteins to fold into unique three-dimensional structures that allows them to function and carry out the instructions of the genome. Thus, comprehending the rules that relate amino acid sequence to structure is fundamental to an understanding of biological processes. Because an amino acid sequence contains all of the information necessary to determine the structure of a protein (1), it should be possible to predict structure from sequence, and subsequently to infer detailed aspects of function from the structure. However, both problems are extremely complex, and it seems unlikely that either will be solved in an exact manner in the near future. It may be possible to obtain approximate solutions by using experimental data to simplify the problem. In this article, we describe how an analysis of allowed amino acid substitutions in proteins can be used to reduce the complexity of sequences and reveal important aspects of structure and function.

## Methods for Studying Tolerance to Sequence Variation

There are two main approaches to studying the tolerance of an amino acid sequence to change. The first method relies on the process of evolution, in which mutations are either accepted or rejected by natural selection. This method has been extremely powerful for proteins such as the globins or cytochromes, for which sequences from many different species are known (2-7). The second approach uses genetic methods to introduce amino acid changes at

specific positions in a cloned gene and uses selections or screens to identify functional sequences. This approach has been used to great advantage for proteins that can be expressed in bacteria or yeast, where the appropriate genetic manipulations are possible (3, 8-11). The end results of both methods are lists of active sequences that can be compared and analyzed to identify sequence features that are essential for folding or function. If a particular property of a side chain, such as charge or size, is important at a given position, only side chains that have the required property will be allowed. Conversely, if the chemical identity of the side chain is unimportant, then many different substitutions will be permitted.

Studies in which these methods were used have revealed that proteins are surprisingly tolerant of amino acid substitutions (2-4, 11). For example, in studying the effects of approximately 1500 single amino acid substitutions at 142 positions in *lac* repressor, Miller and co-workers found that about one-half of all substitutions were phenotypically silent (11). At some positions, many different, nonconservative substitutions were allowed. Such residue positions play little or no role in structure and function. At other positions, no substitutions or only conservative substitutions were allowed. These residues are the most important for *lac* repressor activity.

What roles do invariant and conserved side chains play in proteins? Residues that are directly involved in protein functions such as binding or catalysis will certainly be among the most conserved. For example, replacing the Asp in the catalytic triad of trypsin with Asn results in a  $10^4$ -fold reduction in activity (12). A similar loss of activity occurs in  $\lambda$  repressor when a DNA binding residue is changed from Asn to Asp (13). To carry out their function, however, these catalytic residues and binding residues must be precisely oriented in three dimensions. Consequently, mutations in residues that are required for structure formation or stability can also have dramatic effects on activity (10, 14-16). Hence, many of the residues that are conserved in sets of related sequences play structural roles.

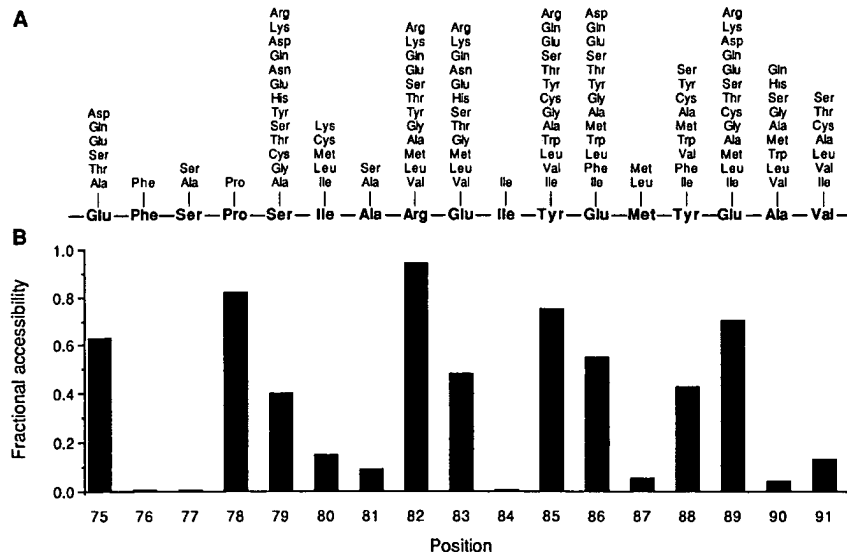
## Substitutions at Surface and Buried Positions

In their initial comparisons of the globin sequences, Perutz and co-workers found that most buried residues require nonpolar side chains, whereas few features of surface side chains are generally conserved (6). Similar results have been seen for a number of protein families (2, 4, 5, 7, 17, 18). An example of the sequence tolerance at surface versus buried sites can be seen in Fig. 1, which shows the allowed substitutions in  $\lambda$  repressor at residue positions that are near the dimer interface but distant from the DNA binding surface of the protein (9). These substitutions were identified by a functional

The authors are in the Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02139.

\*Present address: Department of Chemistry and Biochemistry and the Molecular Biology Institute, University of California, Los Angeles, Los Angeles, CA 90024.

**Fig. 1. (A)** Amino acid substitutions allowed in a short region of  $\lambda$  repressor. The wild-type sequence is shown along the center line. The allowed substitutions shown above each position were identified by randomly mutating one to three codons at a time by using a cassette method and applying a functional selection (9). **(B)** The fractional solvent accessibility (42) of the wild-type side chain in the protein dimer (43) relative to the same atoms in an Ala-X-Ala model tripeptide.



selection after cassette mutagenesis. A histogram of side chain solvent accessibility in the crystal structure of the dimer is also shown in Fig. 1. At six positions, only the wild-type residue or relatively conservative substitutions are allowed. Five of these positions are buried in the protein. In contrast, most of the highly exposed positions tolerate a wide range of chemically different side chains, including hydrophilic and hydrophobic residues. Hence, it seems that most of the structural information in this region of the protein is carried by the residues that are solvent inaccessible.

## Constraints on Core Sequences

Because core residue positions appear to be extremely important for protein folding or stability, we must understand the factors that dictate whether a given core sequence will be acceptable. In general, only hydrophobic or neutral residues are tolerated at buried sites in proteins, undoubtedly because of the large favorable contribution of the hydrophobic effect to protein stability (19). For example, Fig. 2 shows the results of genetic studies used to investigate the substitutions allowed at residue positions that form the hydrophobic core of the  $\text{NH}_2$ -terminal domain of  $\lambda$  repressor (20). The acceptable core sequences are composed almost exclusively of Ala, Cys, Thr, Val, Ile, Leu, Met, and Phe. The acceptability of many different residues at each core position presumably reflects the fact that the hydrophobic effect, unlike hydrogen bonding, does not depend on specific residue pairings. Although it is possible to imagine a hypothetical core structure that is stabilized exclusively by residues forming hydrogen bonds and salt bridges, such a core would probably be difficult to construct because hydrogen bonds require pairing of donors and acceptors in an exact geometry. Thus the repertoire of possible structures that use a polar core would probably be extremely limited (21). Polar and charged residues are occasionally found in the cores of proteins, but only at positions where their hydrogen bonding needs can be satisfied (22).

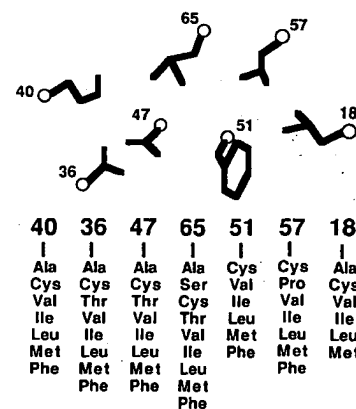
The cores of most proteins are quite closely packed (23), but some volume changes are acceptable. In  $\lambda$  repressor, the overall core volume of acceptable sequences can vary by about 10%. Changes at individual sites, however, can be considerably larger. For example, as shown in Fig. 2, both Phe and Ala are allowed at the same core position in the appropriate sequence contexts. Large volume changes at individual buried sites have also been observed in

phylogenetic studies, where it has been noted that the size decreases and increases at interacting residues are not necessarily related in a simple complementary fashion (5, 7, 17). Rather, local volume changes are accommodated by conformational changes in nearby side chains and by a variety of backbone movements.

## The Informational Importance of the Core

With occasional exceptions, the core must remain hydrophobic and maintain a reasonable packing density. However, since the core is composed of side chains that can assume only a limited number of conformations (24), efficient packing must be maintained without steric clashes. How important are hydrophobicity, volume, and steric complementarity in determining whether a given sequence can form an acceptable core? Each factor is essential in a physical sense, as a stable core is probably unable to tolerate unsatisfied hydrogen bonding groups, large holes, or steric overlaps (25). However, in an informational sense, these factors are not equivalent. For example, in experiments in which three core residues of  $\lambda$  repressor were mutated simultaneously, volume was a relatively unimportant informational constraint because three-quarters of all possible combinations of the 20 naturally occurring amino acids had volumes within the range tolerated in the core, and yet most of these sequences were unacceptable (20). In contrast, of the sequences that contained only

**Fig. 2.** Amino acid substitutions allowed in the core of  $\lambda$  repressor. The wild-type side chains are shown pictorially in the approximate orientation seen in the crystal structure (43). The lists of allowed substitutions at each position are shown below the wild-type side chains. These substitutions were identified by randomly mutating one to four residues at a time by using a cassette method and applying a functional selection (20). Not all substitutions are allowed in every sequence background.



the appropriate hydrophobic residues, a significant fraction were acceptable. Hence, the hydrophobicity of a sequence contains more information about its potential acceptability in the core than does the total side chain volume. Steric compatibility was intermediate between volume and hydrophobicity in informational importance.

## The Informational Importance of Surface Sites

We have noted that many surface sites can tolerate a wide variety of side chains, including hydrophilic and hydrophobic residues. This result might be taken to indicate that surface positions contain little structural information. However, Bashford *et al.*, in an extensive analysis of globin sequences (4), found a strong bias against large hydrophobic residues at many surface positions. At one level, this may reflect constraints imposed by protein solubility, because large patches of hydrophobic surface residues would presumably lead to aggregation. At a more fundamental level, protein folding requires a partitioning between surface and buried positions. Consequently, to achieve a unique native state without significant competition from other conformations, it may be important that some sites have a decided preference for exterior rather than interior positions. As a result, many surface sites can accept hydrophobic residues individually, but the surface as a whole can probably tolerate only a moderate number of hydrophobic side chains.

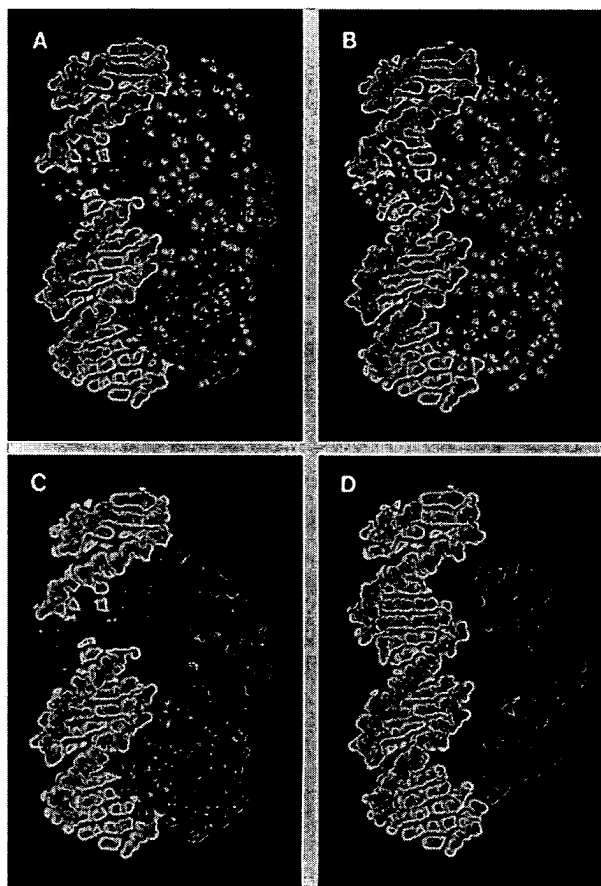
## Identification of Residue Roles from Sets of Sequences

Often, a protein of interest is a member of a family of related sequences. What can we infer from the pattern of allowed substitutions at positions in sets of aligned sequences generated by genetic or phylogenetic methods? Residue positions that can accept a number of different side chains, including charged and highly polar residues, are almost certain to be on the protein surface. Residue positions that remain hydrophobic, whether variable or not, are likely to be buried within the structure. In Fig. 3, those residue positions in  $\lambda$  repressor that can accept hydrophilic side chains are shown in orange and those that cannot accept hydrophilic side chains are shown in green. The obligate hydrophobic positions define the core of the structure, whereas positions that can accept hydrophilic side chains define the surface.

Functionally important residues should be conserved in sets of active sequences, but it is not possible to decide whether a side chain is functionally or structurally important just because it is invariant or conserved. To make this distinction requires an independent assay of protein folding. The ability of a mutant protein to maintain a stably folded structure can often be measured by biophysical techniques, by susceptibility to intracellular proteolysis (26), or by binding to antibodies specific for the native structure (27, 28). In the latter cases, it is possible to screen proteins in mutated clones for the ability to fold even if these proteins are inactive. Sets of sequences that allow formation of a stable structure can then be compared to the sets that allow both folding and function, with the active site or binding residues being those that are variable in the set of stable proteins but invariant in the set of functional proteins. The DNA-binding residues of Arc repressor were identified by this method (8). The receptor-binding residues of human growth hormone were also identified by comparing the stabilities and activities of a set of mutant sequences (28). However, in this case, the mutants were generated as hybrid sequences between growth hormone and related hormones with different binding specificities.

## Implications for Structure Prediction

At present, the only reliable method for predicting a low-resolution tertiary structure of a new protein is by identifying sequence similarity to a protein whose structure is already known (29, 30). However, it is often difficult to align sequences as the level of sequence similarity decreases, and it is sometimes impossible to detect statistically significant sequence similarity between distantly related proteins. Because the number of known sequences is far greater than the number of known structures, it would be advantageous to increase the reach of the available structural information by improving methods for detecting distant sequence relations and for subsequently aligning these sequences based on structural principles. In a normal homology search, the sequence database is scanned with a single test sequence, and every residue must be weighted equally. However, some residues are more important than others and should be weighted accordingly. Moreover, certain regions of the protein are more likely to contain gaps than others. Both kinds of information can be obtained from sequence sets, and several techniques have



**Fig. 3.** Tolerance of positions in the  $\text{NH}_2$ -terminal domain of  $\lambda$  repressor to hydrophilic side chains. The complex (43) of the repressor dimer (blue) and operator DNA (white) is shown. In (A), positions that can tolerate hydrophilic side chains are shown in orange. The same side chains are shown in (B) without the remaining protein atoms. In (C), positions that require hydrophobic or neutral side chains are shown in green. These side chains are shown in (D) without the remaining protein atoms. About three-fourths of the 92 side chains in the  $\text{NH}_2$ -terminal domain are included in both (B) and (D). The remaining positions have not been tested. Data are from (9, 14, 20, 27, 44).

been used to combine such information into more appropriately weighted sequence searches and alignments (31). These methods were used to align the sequences of retroviral proteases with aspartic proteases, which in turn allowed construction of a three-dimensional model for the protease of human immunodeficiency virus type 1 (29). Comparison with the recently determined crystal structure of this protein revealed reasonable agreement in many areas of the predicted structure (32).

The structural information at most surface sites is highly degenerate. Except for functionally important residues, exterior positions seem to be important chiefly in maintaining a reasonably polar surface. The information contained in buried residues is also degenerate, the main requirement being that these residues remain hydrophobic. Thus, at its most basic level, the key structural message in an amino acid sequence may reside in its specific pattern of hydrophobic and hydrophilic residues. This is meant in an informational sense. Clearly, the precise structure and stability of a protein depends on a large number of detailed interactions. It is possible, however, that structural prediction at a more primitive level can be accomplished by concentrating on the most basic informational aspects of an amino acid sequence. For example, amphipathic patterns can be extracted from aligned sets of sequences and used, in some cases, to identify secondary structures.

If a region of secondary structure is packed against the hydrophobic core, a pattern of hydrophobic residues reflecting the periodicity of the secondary structure is expected (33, 34). These patterns can be obscured in individual sequences by hydrophobic residues on the protein surface. It is rare, however, for a surface position to remain hydrophobic over the course of evolution. Consequently, the amphipathic patterns expected for simple secondary structures can be much clearer in a set of related sequences (6). This principle is illustrated in Fig. 4, which shows helical hydrophobic moment plots for the Antennapedia homeodomain sequence (Fig. 4A) and for a composite sequence derived from a set of homologous homeodomain proteins (Fig. 4B) (35). The hydrophobic moment is a simple measure of the degree of amphipathic character of a sequence in a given secondary structure (34). The amphipathic character of the three  $\alpha$ -helical regions in the Antennapedia protein (36) is clearly revealed only by the analysis of the combined set of homeodomain sequences. The secondary structure of Arc repressor, a small DNA-binding protein, was recently predicted by a similar method (8) and confirmed by nuclear magnetic resonance studies (37).

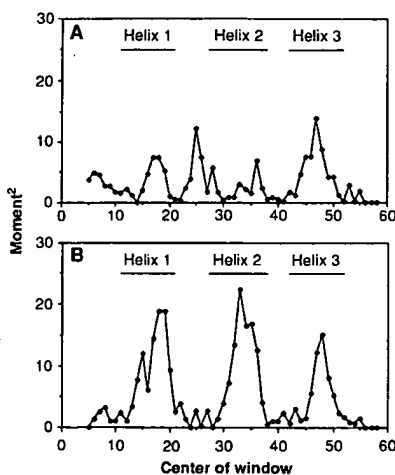
The specific pattern of hydrophobic and hydrophilic residues in an amino acid sequence must limit the number of different structures a given sequence can adopt and may indeed define its overall fold. If this is true, then the arrangement of hydrophobic and hydrophilic residues should be a characteristic feature of a particular fold. Sweet and Eisenberg have shown that the correlation of the pattern of hydrophobicity between two protein sequences is a good criterion for their structural relatedness (38). In addition, several studies indicate that patterns of obligatory hydrophobic positions identified from aligned sequences are distinctive features of sequences that adopt the same structure (4, 29, 38, 39). Thus, the order of hydrophobic and hydrophilic residues in a sequence may actually be sufficient information to determine the basic folding pattern of a protein sequence.

Although the pattern of sequence hydrophobicity may be a characteristic feature of a particular fold, it is not yet clear how such patterns could be used for prediction of structure *de novo*. It is important to understand how patterns in sequence space can be related to structures in conformation space. Lau and Dill have approached this problem by studying the properties of simple sequences composed only of H (hydrophobic) and P (polar) groups on two-dimensional lattices (40). An example of such a representa-

tion is shown in Fig. 5. Residues adjacent in the sequence must occupy adjacent squares on the lattice, and two residues cannot occupy the same space. Free energies of particular conformations are evaluated with a single term, an attraction of H groups. By considering chains of ten residues, an exhaustive conformational search for all 1024 possible sequences of H and P residues was possible. For longer sequences only a representative fraction of the allowed sequence or conformation space could be explored. The significant results were as follows: (i) not all sequences can fold into a "native" structure and only a few sequences form a unique native structure; (ii) the probability that a sequence will adopt a unique native structure increases with chain length; and (iii) the native states are compact, contain a hydrophobic core surrounded by polar residues, and contain significant secondary structure. Although the gap between these two-dimensional simulations and three-dimensional structures is large, the use of simple rules and sequence representations yields results similar to those expected for real proteins. Three-dimensional lattice methods are also beginning to be developed and evaluated (41).

## Summary

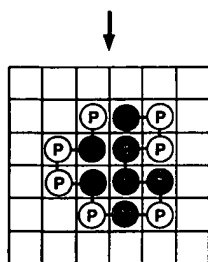
There is more information in a set of related sequences than in a single sequence. A number of practical applications arise from an analysis of the tolerance of residue positions to change. First, such information permits the evaluation of a residue's importance to the function and stability of a protein. This ability to identify the essential elements of a protein sequence may improve our understanding of the determinants of protein folding and stability as well as protein function. Second, patterns of tolerance to amino acid substitutions of varying hydrophilicity can help to identify residues likely to be buried in a protein structure and those likely to occupy



**Fig. 4.** Helical hydrophobic moments calculated by using (A) the Antennapedia homeodomain sequence or (B) a set of 39 aligned homeodomain sequences (35). The bars indicate the extent of the helical regions identified in nuclear magnetic resonance studies of the Antennapedia homeodomain (36). To determine hydrophobic moments, residues were assigned to one of three groups: H1 (high hydrophobicity = Trp, Ile, Phe, Leu, Met, Val, or Cys); H2 (medium hydrophobicity = Tyr, Pro, Ala, Thr,

His, Gly, or Ser); and H3 (low hydrophobicity = Gln, Asn, Glu, Asp, Lys, or Arg). For the aligned homeodomain sequences, the residues at each position were sorted by their hydrophobicity by using the scale of Fauchere and Pliska (45). Arg and Lys were not counted unless no other residue was found at the position, because they contain long aliphatic side chains and can thereby substitute for nonpolar residues at some buried sites. To account for possible sequence errors and rare exceptions, the most hydrophilic residue allowed at each position was discarded unless it was observed twice. The second most hydrophilic residue was then chosen to represent the hydrophobicity of each position. An eight-residue window was used and the vectors projected radially every 100°. The vector magnitudes were assigned a value of 1, 0, or -1 for positions where the hydrophobicity group was H1, H2, or H3, respectively.

P H P P H P H P H H P P H



**Fig. 5.** A representation of one compact conformation for a particular sequence of H and P residues on a two-dimensional square lattice. [Adapted from (40), with permission of the American Chemical Society]

surface positions. The amphipathic patterns that emerge can be used to identify probable regions of secondary structure. Third, incorporating a knowledge of allowed substitutions can improve the ability to detect and align distantly related proteins because the essential residues can be given prominence in the alignment scoring.

As more sequences are determined, it becomes increasingly likely that a protein of interest is a member of a family of related sequences. If this is not the case, it is now possible to use genetic methods to generate lists of allowed amino acid substitutions. Consequently, at least in the short term, it may not be necessary to solve the folding problem for individual protein sequences. Instead, information from sequence sets could be used. Perhaps by simplifying sequence space through the identification of key residues, and by simplifying conformation space as in the lattice methods, it will be possible to develop algorithms to generate a limited number of trial structures. These trial structures could then, in turn, be evaluated by further experiments and more sophisticated energy calculations.

#### REFERENCES AND NOTES

1. C. J. Epstein, R. F. Goldberger, C. B. Anfinsen, *Cold Spring Harbor Symp. Quant. Biol.* **28**, 439 (1963); C. B. Anfinsen, *Science* **181**, 223 (1973).
2. R. E. Dickerson, *Sci. Am.* **242**, 136 (March 1980).
3. M. D. Hampsey, G. Das, F. Sherman, *FEBS Lett.* **231**, 275 (1988).
4. D. Bashford, C. Chothia, A. M. Lesk, *J. Mol. Biol.* **196**, 199 (1987).
5. A. M. Lesk and C. Chothia, *ibid.* **136**, 225 (1980).
6. M. F. Perutz, J. C. Kendrew, H. C. Watson, *ibid.* **13**, 669 (1965).
7. C. Chothia and A. M. Lesk, *Cold Spring Harbor Symp. Quant. Biol.* **52**, 399 (1987).
8. J. U. Bowie and R. T. Sauer, *Proc. Natl. Acad. Sci. U.S.A.* **86**, 2152 (1989).
9. J. F. Reidhaar-Olson and R. T. Sauer, *Science* **241**, 53 (1988); *Proteins Struct. Funct. Genet.*, in press.
10. D. Shortle, *J. Biol. Chem.* **264**, 5315 (1989).
11. J. H. Miller et al., *J. Mol. Biol.* **131**, 191 (1979).

12. S. Sprang et al., *Science* **237**, 905 (1987); C. S. Craik, S. Roczniak, C. Largman, W. J. Rutter, *ibid.*, p. 909.
13. H. C. M. Nelson and R. T. Sauer, *J. Mol. Biol.* **192**, 27 (1986).
14. M. H. Hecht, J. M. Sturtevant, R. T. Sauer, *Proc. Natl. Acad. Sci. U.S.A.* **81**, 5685 (1984).
15. T. Alber, D. Sun, J. A. Nye, D. C. Muchmore, B. W. Matthews, *Biochemistry* **26**, 3754 (1987).
16. D. Shortle and A. K. Meeker, *Proteins Struct. Funct. Genet.* **1**, 81 (1986).
17. A. M. Lesk and C. Chothia, *J. Mol. Biol.* **160**, 325 (1982).
18. W. R. Taylor, *ibid.* **188**, 233 (1986).
19. W. Kauzmann, *Adv. Protein Chem.* **14**, 1 (1959); R. L. Baldwin, *Proc. Natl. Acad. Sci. U.S.A.* **83**, 8069 (1986).
20. W. A. Lim and R. T. Sauer, *Nature* **339**, 31 (1989); in preparation.
21. Lesk and Chothia (5) have argued that a protein core composed solely of hydrogen-bonded residues would also be inviable on evolutionary grounds, as a mutational change in one core residue would require compensating changes in any interacting residue or residues to maintain a stable structure.
22. T. M. Gray and B. W. Matthews, *J. Mol. Biol.* **175**, 75 (1984); E. N. Baker and R. E. Hubbard, *Prog. Biophys. Mol. Biol.* **44**, 97 (1984).
23. F. M. Richards, *J. Mol. Biol.* **82**, 1 (1974).
24. J. W. Ponder and F. M. Richards, *ibid.* **193**, 775 (1987).
25. J. T. Kellis, Jr., K. Nyberg, A. R. Fersht, *Biochemistry* **28**, 4914 (1989); W. S. Sandberg and T. C. Terwilliger, *Science* **245**, 54 (1989).
26. A. A. Pakula and R. T. Sauer, *Proteins Struct. Funct. Genet.* **5**, 202 (1989).
27. B. C. Cunningham and J. A. Wells, *Science* **244**, 1081 (1989); R. M. Breyer and R. T. Sauer, *J. Biol. Chem.* **264**, 13348 (1989).
28. B. C. Cunningham, P. Jhurani, P. Ng, J. A. Wells, *Science* **243**, 1330 (1989).
29. L. H. Pearl and W. R. Taylor, *Nature* **329**, 351 (1987).
30. W. J. Brown et al., *J. Mol. Biol.* **42**, 65 (1969); J. Greer, *ibid.* **153**, 1027 (1981); J. M. Berg, *Proc. Natl. Acad. Sci. U.S.A.* **85**, 99 (1988).
31. W. R. Taylor, *Protein Eng.* **2**, 77 (1988).
32. M. A. Navia et al., *Nature* **337**, 615 (1989).
33. M. Schiffer and A. B. Edmundson, *Biophys. J.* **7**, 121 (1967); V. I. Lim, *J. Mol. Biol.* **88**, 857 (1974); *ibid.*, p. 873.
34. D. Eisenberg, R. M. Weiss, T. C. Terwilliger, *Nature* **299**, 371 (1982); D. Eisenberg, D. Schwarz, M. Komaromy, R. Wall, *J. Mol. Biol.* **179**, 125 (1984); D. Eisenberg, R. M. Weiss, T. C. Terwilliger, *Proc. Natl. Acad. Sci. U.S.A.* **81**, 140 (1984).
35. T. R. Burglin, *Cell* **53**, 339 (1988).
36. G. Otting et al., *EMBO J.* **7**, 4305 (1988).
37. J. N. Breg, R. Boelens, A. V. E. George, R. Kaptein, *Biochemistry* **28**, 9826 (1989); M. G. Zagorski, J. U. Bowie, A. K. Vershon, R. T. Sauer, D. J. Patel, *ibid.*, p. 9813.
38. R. M. Sweet and D. Eisenberg, *J. Mol. Biol.* **171**, 479 (1983).
39. J. U. Bowie, N. D. Clarke, C. O. Pabo, R. T. Sauer, *Proteins Struct. Funct. Genet.*, in preparation.
40. K. F. Lau and K. A. Dill, *Macromolecules* **22**, 3986 (1989).
41. A. Sikorski and J. Skolnick, *Proc. Natl. Acad. Sci. U.S.A.* **86**, 2668 (1989); A. Kolinski, J. Skolnick, R. Yaris, *Biopolymers* **26**, 937 (1987); D. G. Covell and R. L. Jernigan, *Biochemistry*, in press.
42. B. Lee and F. M. Richards, *J. Mol. Biol.* **55**, 379 (1971).
43. S. R. Jordan and C. O. Pabo, *Science* **242**, 893 (1988).
44. R. M. Breyer, thesis, Massachusetts Institute of Technology, Cambridge (1988).
45. J.-L. Fauchere and V. Pliska, *Eur. J. Med. Chem.-Chim. Ther.* **18**, 369 (1983).
46. We thank C. O. Pabo and S. Jordan for coordinates of the NH<sub>2</sub>-terminal domain of  $\lambda$  repressor and its operator complex. We also thank P. Schimmel for the use of his graphics system and J. Burnbaum and C. Francklyn for assistance. Supported in part by NIH grant AI-15706 and predoctoral grants from NSF (J.R.-O.) and Howard Hughes Medical Institute (W.A.L.).

**This Page is Inserted by IFW Indexing and Scanning  
Operations and is not part of the Official Record**

**BEST AVAILABLE IMAGES**

Defective images within this document are accurate representations of the original documents submitted by the applicant.

Defects in the images include but are not limited to the items checked:

- ☐ **BLACK BORDERS**
- ☐ **IMAGE CUT OFF AT TOP, BOTTOM OR SIDES**
- ☐ **FADED TEXT OR DRAWING**
- ☐ **BLURRED OR ILLEGIBLE TEXT OR DRAWING**
- ☐ **SKEWED/SLANTED IMAGES**
- ☐ **COLOR OR BLACK AND WHITE PHOTOGRAPHS**
- ☐ **GRAY SCALE DOCUMENTS**
- ☐ **LINES OR MARKS ON ORIGINAL DOCUMENT**
- ☐ **REFERENCE(S) OR EXHIBIT(S) SUBMITTED ARE POOR QUALITY**
- ☐ **OTHER:** \_\_\_\_\_

**IMAGES ARE BEST AVAILABLE COPY.**

**As rescanning these documents will not correct the image problems checked, please do not report these problems to the IFW Image Problem Mailbox.**